Utilizing constituent structure for compound analysis

Jón Friðrik Daðason & Kristín Bjarnadóttir
The Árni Magnússon Institute for Icelandic Studies
University of Iceland
E-mail: jfd1@hi.is, kristinb@hi.is

Abstract

Compounding is extremely productive in Icelandic and multi-word compounds are common. The likelihood of finding previously unseen compounds in texts is thus very high, which makes out-of-vocabulary words a problem in the use of NLP tools. The tool described in this paper splits Icelandic compounds and shows their binary constituent structure. The probability of a constituent in an unknown (or unanalysed) compound forming a combined constituent with either of its neighbours is estimated, with the use of data on the constituent structure of over 240 thousand compounds from the Database of Modern Icelandic Inflection, and word frequencies from Íslenskur orðasjóður, a corpus of approx. 550 million words. Thus, the structure of an unknown compound is derived by comparison with compounds with partially the same constituents and similar structure in the training data. The granularity of the split returned by the decompounder is important in tasks such as semantic analysis or machine translation, where a flat (non-structured) sequence of constituents is insufficient.

Keywords: decompounding, constituent structure, Icelandic compounds

1. Introduction

Compounding is extremely productive in Icelandic, which poses problems in certain NLP tasks, as the result of the productivity is a quantity of unknown words. Many NLP tasks, including part-of-speech tagging, machine translation and information retrieval, may rely on lexicons with a good coverage of the vocabulary, and the tasks can be adversely affected by the presence of out-of-vocabulary words. The success of NLP tools for Icelandic can therefore be greatly enhanced by compound splitting or decompounding, i.e., the process of breaking compounds into constituent parts.

Decompounding of unknown words has proven useful for other languages, for tasks such as machine translation (Brown, 2002; Koehn and Knight, 2003; Alfonseca, 2008), information retrieval (Hedlund et al., 2001; Braschler et al., 2003), and speech recognition (Adda-Decker et al., 2000). The difference between the methods in general use and the method proposed in this paper is that here the constituent structure of the compounds is used to analyse the unknown parts.

Assuming binary branching (Bjarnadóttir, 2005), each compound is split into two parts, i.e., modifier as a first part, and head as a second part. In Icelandic, the second part of a compound is always the morphological (i.e., inflectional) head. Compounds can be formed by joining any combination of the open word classes, although noun-noun compounding is by far the most productive, and will be used for demonstration in this paper. However, the decompounder described here works equally well for other combinations of the open word classes.

In this paper, a method for generating the constituent structures of compound words is presented. The method is based on the probability that pairs of modifiers and heads can form a compound together, as derived from a large corpus of manually annotated compounds. Using the training data, it is possible to estimate the probability of unknown compounds by comparison with known compounds with a similar structure. The constituent structures can be used in order to split unknown compounds at various levels of granularity, depending on the task at hand. This method is then evaluated on a set of manually annotated Icelandic compounds.

The paper is structured as follows. Section 2 contains a description of Icelandic compounds, and previous work is described in section 3. The description of the decompounder is the body of the paper, with methodology in Section 4 and evaluation in Section 5. Section 6 contains the conclusion and thoughts on future work.

2. Compounding in Icelandic

An example of an Icelandic noun-noun compound is *skólabókasafn* ‘school library’ (*skóla* ‘school’ + *bókasafn* ‘library’), where the second part, *bókasafn* ‘library’, is also a compound (*bóka* ‘book’ + *safn* ‘collection’). The structure of a compound can be ambiguous, as is the case in the above example; the word *skólabókasafn* could potentially also refer to a collection of textbooks (*skólabóka* ‘school books’ + *safn* ‘collection’).

As seen in the example above, the rules of compounding are recursive, as modifiers and heads may be compounds themselves. There is no theoretical limit to the recursivity of compounding in Icelandic, but in reality words with more than six constituents are rare. One such example is *Alþjóðadýraheilbrigðismálastofnun* ‘World Organization for Animal Health’, with the constituents *Al* ‘All’, *þjóða* ‘nations’, *dýra* ‘animals’, *heil+brigðis* ‘health’ (lexicalized compound), *mála* ‘matters’, and *stofnun* ‘organization’, as shown in Fig. 1.

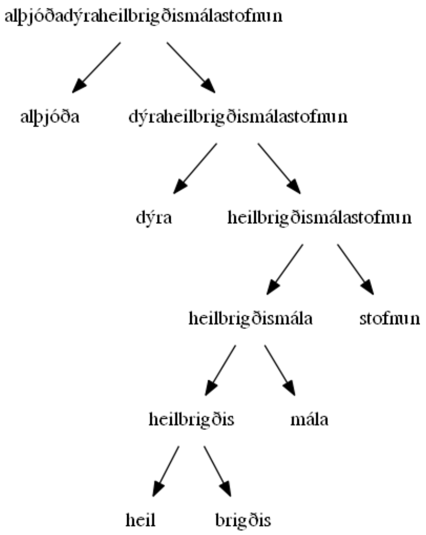

Fig. 1. The constituent structure of *Alþjóðaheilbrigðismálastofnun*.

The analysis of Icelandic compounds is further complicated by variation in combining forms. Nominal modifiers can thus appear as stems or inflectional forms, i.e., in the genitive, singular or plural, or (rarely) the dative, and link phonemes also occur (Bjarnadóttir, 2002). The choice between the first three of these seems to be arbitrary, but not free, i.e., the form itself can be said to be lexicalized (Bjarnadóttir, 1995), as in Table 1:

Stem	Gen.sg.	Gen.pl.	Meaning
bóksala	**bókarsala*	**bókasala*	‘book store’
**bókkápa*	*bókarkápa*	**bókakápa*	‘book cover’
**bókbúð*	**bókarbúð*	*bókabúð*	‘book store’

Table 1. Lexicalization of form in the first part of noun-noun compounds.

The choice between variant forms is usually not meaning-related, as can be seen in the words *barnsmeðlag* and *barnalífeyrir* ‘child support/child maintenance/child allowance’ (*barn,* gen.sg. *barns,* gen.pl. *barna,* ‘child’, *meðlag/lífeyrir* ‘allowance/support/pension’). The word *barnsmeðlag* is used of child support paid by a parent, but *barnalífeyrir* refers to payment by an official body, e.g., the government, etc. The point is that the choice of genitive singular or genitive plural in the modifier does not have a semantic significance; both words can apply to benefits due to one or more children.[1] The variant forms of nouns allowed in the first part of compounds is thus determined by convention, independently of lexicalized semantic relations between the constituent parts.

As inflectional word forms in Icelandic are highly ambiguous (Bjarnadóttir, 2012), lemmatization of the constituents is needed for the disambiguation of the modifiers. The word *andahyggjumaður* is a case in point, as the genitive plural *anda* can be lemmatized as either *andi* ‘spirit’ or *önd* ‘duck’. The other two base words in the compound are unambiguous, i.e., *hyggja* ‘thought’ and *maður* ‘man’. The meaning of the compound could therefore be either ‘spiritualist’ or ‘duck-minded person’ (as of someone specializing in ducks). The first meaning is the correct one, i.e., the accepted or lexicalized version, with the first part *andahyggja* ‘spiritualism’, although the second reading is also in accordance with the rules of productive compound formation. It should be noted that there is no structural morphological difference between productive compound formation and lexicalized compounds in Icelandic, i.e., they are formally compositional to the same degree.[2] The structure of all compounds can therefore be analysed by the same decompounder, but the lemmatization of ambiguous constituents entails the use of semantic features, as in the differentiation of ‘spirit’ and ‘duck’ in *anda*.

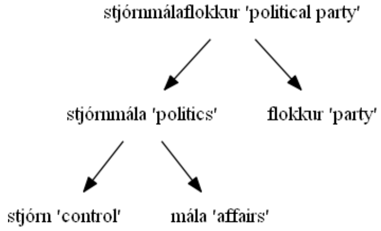

Fig. 2. The constituent structure of *stjórnmálaflokkur.*

The granularity of the split returned by the decompounder is important, as indicated by the examples *andahyggjumaður* and *stjórnmálaflokkur* above. For certain NLP tasks, such as PoS-tagging and lemmatization (of syntactic atoms, i.e., words), it might be sufficient to split compounds into constituent parts (without structural analysis), as the morphological head is generally the only part of a compound that is modified by inflection. For this purpose, knowing the full structure of the compound and being able to specify a less granular split is not important. The same might not apply for other tasks, such as semantic analysis or machine translation. This can be demonstrated by the compound *stjórnmálaflokkur* ‘political party’ (*stjórnmála* [*stjórn* ‘control’ + *mál* ‘matters’] = ‘politics’ + *flokkur* ‘party’). Translating the base words that the compound *stjórnmálaflokkur* is made up of would result in a mistranslation (e.g., ‘control affairs party’ instead of ‘political party’). Similarly, incorrectly as-

[1] This is a simplification, as there are compounds where number is significant in the modifier, as in *bróðursonur* (*bróður* gen.sg. ‘brother’ + *sonur* sg. ‘son’) ‘nephew’, i.e. the son of one's brother (pl. *bróðursynir*); *bræðrasynir* (*bræðra* gen.pl. ‘brother’, *synir* pl. ‘sons’) ‘sons of brothers’’ or ‘sons of one’s brothers’’. The form *bræðrasonur* ‘the son of brothers’’ is not found. This kind of distinction of number in the first part of a noun-noun compound is rare.

[2] Archaic forms occur in a few words; these can easily be listed and do not represent a problem.

suming that the word is split as *stjórn* ‘control’ + *málaflokkur* ‘category’ would also result in a mistranslation. Thus, being able to specify splits at various levels of granularity is a very useful feature of a decompounder that is intended for use in a variety of NLP tasks.

3. Related works

Brown (2002) describes a method by which compounds can be broken down into cognates and words found in a translation lexicon between the target (compounding) language and a non-compounding language (in this case, German and English, respectively).

Koehn and Knight (2003) find all possible splits for unknown compounds, where each part is a known word (allowing for inflectional and linking morphemes in between). Each candidate split is scored according to the geometric mean of the word frequencies of its parts. They also use a translation lexicon between the target language and a non-compounding language, giving preference to splits where the compound parts have a one-to-one correspondence to the translation in the other language. This method achieves 94% precision and 90% recall when evaluated on a manually annotated corpus of German compounds.

Schiller (2005) uses weighted finite-state transducers to find possible segmentations for compound words, each of which is weighted as the product of the probability of its parts. The probability of an individual part (which may either be a modifier or a head) is derived from a training corpus of manually annotated compounds. When evaluated on a corpus of German compounds from medical and newspaper texts, this method achieves a precision of 96-98% and a recall of 98%-99%.

Alfonseca et al. (2008) combine a number of different methods in a support vector machine (SVM), achieving significantly improved results over any of the included methods. The authors report a precision of 83% and a recall of 79% when evaluated on a corpus of German web queries.

4. Methodology

The method described here can be summed up in the following steps:

1. A potential compound is split into all possible sequences of base words it could consist of.
2. A constituent structure (a binary tree) is built bottom-up for each possible segmentation of the word. In each step, the two neighbouring parts with the greatest probability of forming a constituent together are joined.
3. The constituent structure with the highest probability is chosen.

4.1 Splitting compounds

The Database of Modern Icelandic Inflection (DMII, Bjarnadóttir, 2012) is a collection of approximately 270.000 Icelandic paradigms, containing approximately 5.8 million inflectional forms, both base words and compounds. Each inflectional form is tagged for word class, lemma, and grammatical features (e.g., gender, number, case, and definiteness for nouns; person, number, tense, etc. for verbs, etc.). In the process of the creation of the DMII, unpublished data on the binary split of each compound has been created. This data is used in the decompounder described here, in the form of a list of over 240.000 compounds, making it possible to construct a binary tree for every compound in the DMII.

Splitting the inflectional forms of the compounds in the DMII into the base words of which they consist, yields a total of approximately 169.000 distinct inflectional forms of base words, of which about 40.000 can appear as a modifier and 146.000 as a head. The difference in numbers stems from the fact that the combining forms of the modifiers are a subset of all inflectional forms, whereas the head can theoretically occur as any inflectional form in its full paradigm.[3]

The head of an Icelandic compound has the same grammatical features as the compound itself. Thus, if a modifier is a compound, then its head is also a potential modifier. Possible segmentations for a compound are therefore any sequence of words that can appear as a modifier, followed by a known head, which the compound could be comprised of.

4.2 Neighbour joining

The probability that two neighbouring parts could be joined to form a combined constituent is estimated from constituent structures that occur in the DMII compound data combined with the frequency of their occurrence in *Íslenskur orðasjóður* (Hallsteinsdóttir et al., 2007), a corpus of approximately 550 million Icelandic words from the web. The probability is calculated as

$$P(mod + head) = \frac{count(mod + head)}{N}$$

where $count(mod + head)$ is the number of compound words within a corpus where these parts appear next to one another and form a compound together, and N is the total number of compounds in the corpus.

The probability of two neighbouring parts forming an unknown compound together is estimated by breaking the resulting constituent structure into smaller parts and multiplying their probabilities. Thus, the probability of *heilsu* ‘health’ + *vara* ‘product’ forming the compound *heilsuvara* ‘health product’ (assuming it were unknown) is estimated by multiplying the probability of *heilsu* appearing as a modifier (to any head) and of *vara* appearing as a head (to any modifier), i.e.,

$$P(heilsu + vara) = \frac{count(heilsu + *)}{N} * \frac{count(* + vara)}{N}$$

where * stands for any word. In a sense, *heilsu*+* could be considered to be a template for any compound with the modifier *heilsu*.

[3] A full paradigm for a noun includes 16 inflectional forms with distinct PoS-tags. The corresponding figure for an adjective is 120, and 106 for a verb. (DMII, http://bin.arnastofnun.is/).

Larger templates are constructed from groups of compounds that are very similar in composition. Training the decompounder on a large collection of constituent structures makes it possible to make certain observations about the structure of compounds that share very similar characteristics. For example, consider *fjármálaráðherra* ‘finance minister’ (*fjármála* ‘finance’ + *ráðherra* ‘minister’), *dómsmálaráðherra* ‘justice minister’ (*dómsmála* ‘justice’ + *ráðherra* ‘minister’) and a number of other similar compounds. They all share the same head, *ráðherra*, and a modifier that is itself a compound with the head *mála* (i.e., *fjár* ‘money’ + *mála* ‘affairs’, *dóms* ‘court’ + *mála* ‘affairs’, etc.). The structure of these compounds can be used to gain an insight into the probability of unknown words with the same structure.

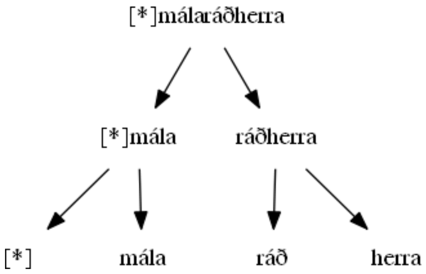

Fig. 3. A template for a group of compounds, such as *fjármálaráðherra* and *dómsmálaráðherra*.

The probability of *samgöngumála* ‘transport’ and *ráðherra* ‘minister’ forming the compound *samgöngumálaráðherra* is estimated by multiplying the probabilities of *samgöngu*+*mála* and **mála*+*ráðherra*.

Constituent structures (binary trees) are built bottom-up for every possible segmentation of the compound, by iteratively joining the two adjacent nodes with the greatest probability of forming a compound together until only one node (the root) remains. The resulting constituent structures are then ranked by the probability of their last joining operation, and the structure with the highest probability is chosen.

In the compound *andahyggjumaður* ‘spiritualist’ (cf. Sec. 2), the probabilities of the adjacent parts suggest the constituent structure *andahyggju*+*maður*.[4] There are in fact a number of terms like *andahyggja* ‘spiritualism’ in use, all of which can be modifiers in compounds ending in *maður* ‘man’, e.g., *dulhyggja* (‘occultism’, from *dulur* ‘secret, hidden’), *efahyggja* (‘skepticism’, from *efi* ‘doubt’), *efnishyggja* (‘materialism’, from *efni* ‘matter, material’), *félagshyggja* (‘socialism, communitarianism’, from *félag* ‘association, society’), *frjálshyggja* (‘libertarianism’, from *frjáls* ‘free’). Disambiguation, and therefore correct lemmatization of the first part *anda* (i.e., *andi* ‘spirit’ and not *önd* ‘duck’), should be possible by semantic clustering at a later stage, but for now the data suffices to produce the correct binary tree, as there is quite a number of compounds with the same constituent structure in the data.

5. Evaluation

The method described in this work is evaluated on a collection of manually annotated Icelandic Wikipedia articles, containing a total of 6.098 words (of which 3.319 are compounds).

The evaluation is two-fold. First, we evaluate the accuracy of the decompounder when used to analyse the structure of compound words. Second, the overall performance of the decompounder is evaluated on the full set of words.

N	Count	Atoms	Binary split	Binary tree
N=2	2.709	99.5%	99.6%	99.5%
N=3	513	93.2%	97.9%	91.8%
N=4+	97	91.8%	96.9%	88.7%
Total	3.319	98.3%	99.2%	98.0%

Table 2. The results of the evaluation on compound word analysis, where N denotes the number of base words the compounds were comprised of.

Table 2 shows the ratio of compounds which the decompounder successfully breaks down into the sequence of base words (‘atoms’) which they are comprised of, the ratio of compounds that were successfully split in two, and finally the ratio of compounds for which the decompounder could correctly determine the entire constituent structure (i.e., their binary tree representation).

The overall performance of the decompounder is evaluated in terms of precision and recall. The precision of the decompounder is computed as the number of correctly split compounds (into binary trees) divided by the total number of words that are split. The recall is computed as the number of correctly split compounds divided by the number of compounds in the text. When evaluated on the full text, the decompounder achieves a precision of 97.6%, a recall of 98.0% and an accuracy of 99.2%.

6. Conclusion and future work

The results presented in this paper show that the method reduces the number of unknown (i.e., unanalyzed) compounds substantially. By using the binary tree of the constituent structure rather than splitting the compounds into flat sequences of words, it is possible to specify splits at various levels of granularity. This is essential for the disambiguation of the compounds, both in the disambiguation of individual constituents (as in *anda* ‘spirit’/‘duck’ in *andahyggjumaður*) and in the disambiguation of the compound structures themselves (as in *stjórnmálaflokkur* ‘political party’).

The method described in this paper will be used in future projects at the Árni Magnússon Institute for Icelandic

[4] A single citation of the compound *hyggjumaður* ‘a man of thought’ in The Written Language Archive at The Árni Magnússon Institute for Icelandic Studies could indicate the other possible binary tree for the word *andahyggjumaður*. However, the citation is from the 16th century, and the word *hyggjumaður* is exceedingly rare.

Studies, such as automatic word excerption for additional vocabulary for the DMII, and for the correction of compounds in the context-sensitive spellchecking application Skrambi (work in progress).

Future work includes exploring the use of semantic clustering in order to obtain better results on unknown compounds. The disambiguation of *anda* in section 4.2 shows an example of this. The word *andi* ‘spirit’ is more convincing in a semantic cluster with the words *dulur* ‘secret’, *efi* ‘doubt’, *efni* ‘matter’, *félag* ‘society’ and *frjáls* ‘free’ than the word *önd* ‘duck’ would be, as any semantic categorization will show. The fact that these words (and quite a few more) are the first constituent in terms for philosophical and political “isms” of all kinds ending in *-hyggja* demonstrates the point.

7. Acknowledgements

The project described in this paper forms a part of a collection of NLP tools being developed at the Árni Magnússon for Icelandic Studies (AMI) in Reykjavík. The prototype was developed in 2012 by Jón Friðrik in his M.Sc. program in Computer Science at the University of Iceland, and thanks are due to his teacher in computational linguistics, Professor Eiríkur Rögnvaldsson. Sigrún Helgadóttir and Professor Sven Þ. Sigurðsson, from the AMI and the University of Iceland respectively, also took an active part in the work on the project, in an advisory capacity; they truly deserve our gratitude. Thanks are also due to Þórdís Úlfarsdóttir, our colleague at the AMI, and anonymous LREC reviewers for their helpful comments, which we have taken into account as far as possible. All mistakes and short-comings are ours, of course.

8. References

Bjarnadóttir, Kristín. (1995). Lexicalization and the Selection of Compounds for a Bilingual Icelandic Dictionary Base. In Ásta Svavarsdóttir, Guðrún Kvaran, Jón Hilmar Jónsson (ritstj.). *Nordiske studier i leksikografi* 3:255-263.

Bjarnadóttir, Kristín (2002). *A Short Description of Icelandic Compounds.* Retrieved from http//www.lexis.hi.is/ kristinb/comp-short.pdf.

Bjarnadóttir, Kristín. (2005). *Afleiðsla og samsetning í generatífri málfræði og greining á íslenskum gögnum.* Reykjavík: Orðabók Háskólans. [Derivation and compounding in generative grammar and the analysis of Icelandic data.]

Bjarnadóttir, Kristín (2012). The Database of Modern Icelandic Inflection. In *Proceedings of the workshop Language Technology for Normalization of Less-Resourced Languages, SaLTMiL 8 - AfLaT, LREC 2012,* pp. 13-18, Istanbul, Turkey.

Braschler, M., Göhring, A., & Schäuble, P. (2003). Eurospider at CLEF 2002. In C. Peters, M. Braschler & J. Gonzalo (Eds.) *Advances in Cross Language Information Retrieval* (Vol. 2785, pp. 164-174): Springer Berlin / Heidelberg.

Brown, R. D. (2002). Corpus-Driven Splitting of Compound Words. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translations* (TMI-2002).

The Database of Modern Icelandic Inflection. [Beygingarlýsing íslensks nútímamáls.] (n.d.) Kristín Bjarnadóttir, editor. The Árni Magnússon Institute for Icelandic Studies. http://bin.arnastofnun.is/

Hallsteinsdóttir, E., Eckart, T., Biemann, C., Quasthoff, U., & Richter, M. (2007). Íslenskur Orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07*, Tartu, Estonia.

Hedlund T., Keskustalo H., Pirkola A., Airio E. & Järvelin K. (2001) Utaclir @ CLEF 2001 – effects of compound splitting and n-gram techniques. In Second *Workshop of the Cross-Language Evaluation Forum (CLEF)*, Revised Papers.

Íslenskur orðasjóður. (n.d.) Universität Leipzig. http://wortschatz.uni-leipzig.de/ws_isl/

Koehn, P., & Knight, K. (2003). Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics* - Volume 1, Budapest, Hungary.

Ritmálssafn Orðabókar Háskólans. (n.d.) The Árni Magnússon Institute for Icelandic Studies. http://www.arnastofnun.is/page/ritmal.

Schiller, A. (2005). German Compound Analysis with wfsc. In *Proceedings of the Fifth International Workshop of Finite State Methods in Natural Language Processing (FSMNLP),* Helsinki, Finland, 239-246.